Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol 3, No 16, 2012
www.iiste.org

Applying the Pythagorean Model to Derive a Correction Factor for Estimating Minimal Competence with Greater Fidelity

Manoj Chakravarty[1*], Waleed H. Al-Bu-Ali[1], Nasir Abdul Latif[2], Marwan Abu-Hijleh[2], Pallab Kumar Ganguly[3]

1. College of Medicine, King Faisal University, Al-Ahsa, Kingdom of Saudi Arabia
2. College of Medicine, Arabian Gulf University, Manama, Kingdom of Bahrain
3. College of Medicine, Al-Faisal University, Riyadh, Kingdom of Saudi Arabia

* E-mail of Corresponding author: mavarty@kfu.edu..bh

Abstract

We propose an alternate psychometric method for judging minimal competence using core knowledge as a major variable. Utilizing the 'Pythagorean' model, we offer to establish a mathematically valid relationship between a borderline candidate's scores in test items reflecting core knowledge, desirable or higher knowledge, and the total raw score obtained in a given test. Using this method, it is possible to establish a mathematical relationship between the above variables and derive a positive integer, named herein as the **Correction Factor** (CF) that would help in providing a better means to identify minimal competence with greater fidelity.

Keywords: Cut-point; Domain-referenced; Minimal competence; Core knowledge, Correction Factor.

1. Introduction

Standards are a systematic way of gathering value judgments, reaching a consensus and applying it to set a definite score to represent a cut-point. Standard setting is a process of deciding "what is good enough" or "how much is enough", and more than two dozen criterion-referenced standard setting methods have been described (Berk, 1986). Many empirical studies claim that different standard setting procedures yield different cut scores (Jaeger, 1989). Emphasis must be made on the understanding that minimum competence does not represent something substandard. Since standards are an expression of values, methods for setting them are systematic ways of gathering value judgments, reaching consensus and expressing that consensus as a single score on a test (Norcini, 2003).

There are two main types of score interpretations; Norm-referenced and Criterion-referenced. A Norm-referenced interpretation involves comparing a person's score with the average score of some relevant group of people. A Criterion-referenced interpretation is made when we compare a person's score with scores that each represent distinct levels of performance in some specific content area or with respect to a behavioural task, (Ebel & Frisbie, 1991).

There is much confusion, even among measurement specialists, about what the term criterion-referenced means. Part of the confusion stems from the fact that "criterion" is used in several other ways by testing specialists. Another part of the confusion relates to the wide variety of interpretations that can be classified correctly as criterion-referenced, (Nitko, 1980).

Hively (1974) and Millman (1974) recognized this ambiguity and offered the term *domain-referenced* as a more exact description of a test designed primarily to optimise absolute or content related interpretations. Domain-referenced test is "any test consisting of a random or stratified random sample of items selected from a well-defined set or class of tasks (a domain)", (Millman, 1974). In this study, we prefer to use the term domain-referenced rather than criterion-referenced for the above reasons.

Existing errors arise due to inappropriateness of any assessment system to measure quality, as existing systems are skewed heavily towards assessment of quantity. Many educational measurement specialists have asserted that the process of establishing pass-fail standards for credentialing purposes is unavoidably arbitrary (Ebel, 1979; Glass, 1978). Available methods of the judgmental or absolute types generally depend on judges' abilities to imagine a minimally competent individual, contributing to high variability between judges. In addition, if judgment criteria are characterized by more than one scale of measurement, the next challenge would be how to combine them into a single pass-fail decision. One could arguably pass someone who in reality may not have been fit to succeed, or could fail to identify someone who in reality may have indeed deserved to succeed. The result of such measurements for determination of minimal competence is a certain numerical cut-point derived either by a fixed pass point score or through the use of judgmental methods that utilize multiple judges to rate test items. To make such a number more meaningful, it is necessary to compare it with something. We should recognize that not all performance standards are points on a scale. Measurement properties should therefore be in

place to address credibility, comprehensiveness, precision, validity and feasibility (Neufield, 1984). Dilemma still exists between how far one could use effective methods to determine cut points to effectively identify minimal competence. Categorization may be achieved by the use of absolute, relative or compromise methods (Livingston & Zeiky, 1982).

The simple professional and ethical solution is to attach an estimate of error to every application for the measurement of standards. Estimation of hypothetical error scores can be measured, and this quantity is called *standard error of measurement* (SEM).

Even though SEM is the most common indicator of the amount of error contained in an observed test-score, its shortcoming is that it provides the same error estimate for everyone in the group. Methods that permit the computation of SEM's of measurement for each of several score ranges have been described (Feldt & Brennan, 1989).

There is no doubt about the empirical evidence available about the extreme vulnerability of any single judge in determining either a stable rank order in concurrent rank-ordering of the same tests or in the great differences in rank- orderings between different judges (Rechter, 1968).

From the above discussion, a plethora of uncertainties are bound to arise. Some of these are:

- Test validation in essence is scientific inquiry into score meaning - nothing more, but also nothing less (Messick, 1989).
- No measure of a single skill can ever be mapped on a non-trivial vision of real success because any problem can be solved in more than one way (Burton, 1978).
- What do tests measure? It is clear that there are no units; the measure is a pure number. Tests have so many independent sources of invalidity that they do not measure anything in particular, nor do they place people in any particular order of anything, except along a single line of 'merit'. (Wilson, 1998).
- Only a perfect score is consistent with the definition of minimal. So to attempt to find an appropriate 'cut-off' score to use as a standard is to engage in a paradox, to indulge in contradiction and to professionalize an absurdity (Berk, 1986).

It is in light of the above concerns that we would like to offer an alternative psychometric approach towards the assessment of minimal competence. The key features of the proposed standard-setting method are:

- Providing an alternative approach towards creating a predictor cut-point with greater reliability and fidelity by using the Pythagorean theorem to interrelate among a set of variables.
- Using core-knowledge as a key factor in determining minimal competence.
- Involving the use of subject-experts as opposed to non-subject experts as judges to identify core (essential) test items.
- Allowing comparison of a minimally competent student's scores in 'core' versus 'desirable' test items and between the student's scores in core test items versus the raw score for a given test.

We are of the opinion that existing psychometric processes may not be the best contributor to judgments regarding estimation of minimal competence. Based upon the above premise, the goals of this study were to:

- Re-examine some fundamental tools of educational measurement.
- Develop a tool for overcoming perceived discrepancies in judging minimal competence.
- Create a better scale for establishing minimal competence by comparing performance in core (essential) with desirable (nice-to-know) and higher-level test items as well as with the total raw score in a given test.
- Provide an alternative ethical and psychometric solution to standard-setting and reduce the prevalence of noise attributable to suppression of error in the categorization of test takers to estimate minimal competence.

2. Method

Anecdotal evidence:

In order to distinguish between test takers who are competent versus those who are not, levels of competencies need to be defined. Cut points that help to arrive at pass-fail decisions, mainly utilize methods that are criterion-referenced. Such methods include absolute or expert (judgmental) methods like those of Nedelsky (1954), Angoff (1971), Ebel (1979), Livingston & Zeiky (1982), compromise method of Hofstee (1980), continuum model of Jaeger (1980), and empirical method of cluster analysis of Sireci et al. (1995). These methods utilize multiple judges who help to determine cut points for identifying minimal competence in a given test. Cutoff

score for pass-fail decisions is established a priori by determining the minimum performance level (MPL). The above methods, in addition to determining cut points, also help to identify borderline students, whose scores could be re-assessed before reaching a final pass-fail decision.

A summary of the above methods is presented for the sake of reference.

- Nedelsky (1954): Oldest procedure used in the health professions for Multiple Choice items using judges to look at each test item and identify the incorrect options that a minimally competent individual would know were wrong.
- Angoff (1971): Judges are used to estimate the probability that a minimally acceptable person would get each test item right.
- Ebel (1979): Judges are asked to rate test items on the basis of two dimensions of relevance and difficulty. For relevancy, three levels consisting of easy, medium and marginal are employed. For difficulty, three levels consisting of easy, medium and hard are employed.
- Livingston & Zeiky (1982): Judges choose individuals who are considered borderline with respect to estimation of minimal competency.
- Hofstee (1980): Judges are asked to specify the maximum required percentage of mastery, minimum required percentage of mastery, maximal acceptable percentage of failures and, minimum acceptable percentage of failures to estimate cut-points.
- Jaeger (1995): Continuum models are used that could either be test-centered or examinee-centered. In test-centered models, judges set the cutoff score by reviewing individual test items and decide on the level of performance in each item. This is used to determine a minimum performance level (MPL) for a given test. In the examinee-centered models, judges determine the cutoff score and make pass-fail decisions about actual examinees after they have written the test.
- Sireci (1997): Cluster method analysis providing cutoff scores based strictly on mathematical criteria by forming two naturally occurring groups based on the minimization of within-group variance and maximization of between-group variance.

Our proposed method aims at improving existing pass-fail decision-making processes by utilizing scores obtained by minimally competent (borderline) candidates in test items that essentially reflect core knowledge. In order to establish a cogent relationship between two sets of variables namely, core test items versus desirable test items, and between core test items versus the total score respectively, we have adopted the 'Pythagorean' model that is both simple as well as appropriate.

The Pythagorean theorem is a relation in Euclidean geometry among the three sides of a right-angled triangle. The theorem can be written as an equation relating the lengths of a triangle with sides **a, b** and **c** often called the Pythagorean equation: $\mathbf{a^2 + b^2 = c^2}$, where **c** represents the length of the hypotenuse and **a** and **b** represent the lengths of the other two sides (**Fig. 1**). When the three integers are positive, the relationship between them is called the Pythagorean triple that represents the lengths of the sides of a right angle triangle where all the three sides have integer lengths (Sally, 2007). In context, the outline of the methodology of the proposed psychometric model is presented as follows:

2.1. Test items reflecting core knowledge are identified by subject matter experts who form a panel of judges from disciplines represented in a given test. Test scores of a candidate in core knowledge test items obtained by this method is compared first, to the candidate's performance in the rest of the test items (desirable or higher level) and second, to the overall score obtained by the candidate in that test. It is thus possible to associate these two relationships using the Pythagorean model and derive a relationship mathematically, named herein as the **Correction Factor** (**CF**) that would be added to the existing cut point obtained through the use of existing traditional psychometric methods. In institutions that have a fixed cut point, this translates into having the **CF** being added to the score obtained by a borderline or minimally competent student. Pass-fail decision based upon use of the proposed method is thus expected to provide a greater degree of fidelity as well as reliability.

2.2. A composite list of test items reflecting core knowledge and higher level or desirable knowledge is created (**Table 1**).

2.3. The total number of core and desirable (nice to know) test items are now plotted to scale represented by the sides **B-C** and **A-C** respectively in the triangle **ABC** (**Fig. 2**).

2.4. The percentage score obtained by the student being assessed for core test items is now plotted against the raw score that is obtained by the student in a given test. (In our hypothetical setting these scores are 70% and 56% respectively) (**Fig. 2**). These scores are plotted along the sides **D-E** and **F-E** respectively in

the triangle **DEF** (**Fig.3**). In order to make the domain-referenced approach more meaningful in context, an analysis is undertaken by comparing the relationship between the score obtained in core test items versus the score obtained in desirable test items, and the score obtained in core test items versus the overall raw score for a given test.

2.5. In order to extrapolate this relationship, we propose to apply the Pythagorean model to derive a proposed mathematical entity named herein as the Correction Factor (**CF**). We propose to apply this method in effectively establishing a mathematical relationship between scores obtained by a minimally competent candidate that is represented by the positive integers related to the score in Core knowledge and Desirable knowledge test items.

2.6. Through our proposed method, the relationship between the total number of core test items and the total number of desirable knowledge test items is derived as the value of the hypotenuse **A-B** in the triangle **ABC** (**Fig. 2**). The relationship between the student's score in core knowledge test items against the raw score obtained in a given test is now derived as the value of the hypotenuse **F-D** in the triangle **DEF** (**Fig.3**). The two variables **A-B** and **F-D** thus derived are then mathematically equated by dividing the value of **F-D** by the value of **A-B** to provide a positive integer called the Correction Factor (**CF**).

2.7. The calculation of the **CF** may be achieved by use of the 'Pythagorean Theorem' where, for a given triangle **abc** (**Fig.1**), the hypotenuse **ab** is calculated as follows:

$$ab^2 = ac^2 + bc^2$$

The proposed **CF** would be derived accordingly, by comparing the magnitudes of the hypotenuses **A-B** (**Fig.2**) and **F-D** (**Fig.3**) mathematically and would be calculated as follows:

$$\text{Correction factor (CF)} = \frac{\text{Calculated value of F-D}}{\text{Calculated value of A-B}}$$

In our hypothetical situation, using the Pythagorean theorem, this would equate as follows:

$$CF = \frac{\sqrt{70^2 + 56^2}}{\sqrt{64^2 + 36^2}} = \frac{\sqrt{8036}}{\sqrt{5392}} = \frac{89.6}{73.4} = 1.22$$

2.8. The final test score obtained by a student in the above hypothetical situation would be the sum of a student's score obtained through existing standard setting procedures and the **CF**, which in our hypothetical case is **1.22**. If a borderline student is then able to equal or better the existing cut point, his/her performance would be considered satisfactory as per our proposed domain-referenced method reflecting a greater degree of fidelity and reliability than existing methods that do not consider performance of a candidate in test items reflecting core knowledge.

3. Discussion

Standards for estimation of cut points in general, appear to be unrelated to the estimation of quality. Standards define things primarily with respect to quantity, and thus while an acceptable standard may not necessarily reflect a true score, a true score in turn, may not necessarily reflect a definitive standard. Effective assessment is a continuous cycle of development, implementation, presentation and evaluation. Medical educators should consider whether the education process is congruent with the students total learning experience. (Fowell, Southgate, Bligh, 1999). Nedelsky, (1954), Angoff (1971), and others, are accredited with methods in the setting of examination standards that are in use despite limitations (Brennen, 1980; Jaeger, 1982). The results of cluster analysis indicate that although the percentage agreement rate is very high, there is nevertheless some disagreement (Violato, et al, 2003). One predominant limitation of available psychometric standards is the high degree of subjectivity linked to judges' competencies. The state of art of standard setting for performance assessment is far from a state of grace (Jaeger, et al, 1996). To offer reprieve from existing limitations, and also to provide an alternative as well as scientific insight towards arriving at cut-points, we intend to offer a more reliable approach to standard setting. We shall find critics with different views of the same situation. This makes us try and find reliability among judges, and by doing so, we might achieve a higher level of intercritic agreement, even if in the process we compromise validity (Eisner, 1988). The method proposed by us, while offering an alternative approach to setting of performance standards, aims at providing greater fidelity. We are of the opinion that our proposed method could offer reduction in anxiety that is attributable to the suppression of

error in categorization of examinees towards pass-fail decisions. Thus, while conceptually providing to normalize scores, we are able to formulate a linear scale and also create a measure that is mathematically valid, simple and user-friendly. Our proposed method takes into consideration the fact that good performance in core test items could be appropriately used to reward a minimally competent candidate. This also ensures that any borderline candidate who has not performed comparatively well in core test items does not benefit from the proposed method, and this helps to support the domain-based approach towards adopting a better decision-making strategy.

4. Conclusion

By use of the proposed method, we offer to provide a more accurate domain-referenced predictor cut-point for judging minimal competence in a given test and to overcome subjectivity issues related to the use of non-subject matter experts to establish predictor cut points on test items reflecting core knowledge. Our proposed method could be used for summative as well as for certification purposes. It could be optimized to work equally well not only in situations where multiple judges are available for determination of minimal competence, but also in curricula that have fixed or absolute standards. We feel that medical schools need to work together through their professional associations, to assure some degree of consistency in applying assessment standards to arrive at better cut-points.

References

Angoff, W. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement*, (pp. 508-600), Washington, DC: American Council on Education.

Berk, R.A. (1986). A consumers guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*; 6:137-72.

Brennen, R.L. (1980). A comparison of the Nedelsky and Angoff citting scores procedures using generalisability theory. *Applied Psychological Measurement*. 4:219-40.

Burton, N. (1978). Societal Standards. *Journal of Educational Measurement*, 15(4), 263-273.

Ebel, R.L. (1979). *Essentials of educational measurement*. Englewood Cliffs, N.J.: Prentice-Hall.

Ebel, R.L., Frisbie, D.A. (1991). *Essentials of Educational Measurement*, 5th Ed, Prentice Hall, Englewood Cliffs, NJ, 2:34-35.

Eisner, E.W. (1988). The primacy of experience and the politics of method. *Educational Researcher*, 17(3), 15-20.

Feldt, L.S. and Brennan, R.L. (1989). Reliability. In R.L. Linn (ed), *Educational Measurement (3rd ed.)*. Washington, DC: American Council on Education.

Fowell, S.L., Southgate, L.J., Bligh, J.G.: Evaluating Assessment: the missing link? *Medical Education*, 1999; 33:276-281.

Glass, G. (1978). Standards and criteria. *Journal of Educational Measurement.* 15, 237-261.

Hively, W. (1974). Introduction to domain-referenced testing. In W. Hively (ed.), *Domain-referenced testing.* (5-15). Englewood Cliffs, NJ: Educational Technology Publications.

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational Measurement*, (3rd. Ed., pp. 485-514). New York Macmillan.

Jaeger, R.M. et al (1996). In, Standard Setting in Student Assessment. *Education Guide No.18,* AMEE, p17.

Livingston S.A, Zeiky M.J. (1982). *Passing scores: a manual for setting standards of performance on educational and occupation tests*. Princeton, New Jersey: Educational Testing Service.

Messick, S. (1989). Meaning and values in test validation. *Educational Researcher*, 18(2), 5-11.

Millman, J. (1974). Program assessment, criterion-referenced tests, and things like that. *Educational Horizons*, 32: 188-192.

Millman, J. (1974). Criterion-referenced measurement. In W.J. Popham (ed.), *Evaluation in education*. Berkeley, CA: McCutchan. p315.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement,* 14: 3-19.

Neufield, V.R. (1984). The design and use of assessment methods for problem-based learning. In Schmidt HG, de Volder ML, eds. *Tutorials in problem-based learning: new directions in training for the health professions. Assen/Maastricht: Van Gorcum.*

Nitko, A.J. (1980). Distinguishing the many varieties of criterion-referenced tests. *Review of educational research*, 50(3), 461-85.

Norcini, J.J. (2003). Setting standards on educational tests. *Medical Education*; 37:464-469.

Rechter, B. & Wilson, N. (1968). Examining for university entrance in Australia: *Current practice. Quarterly Review of Australian Education,* 2(2).

Sally Judith, D., Paul Sally, (2007). Chapter 3: "Pythagorean triples". *Roots to research: a vertical development of mathematical problems.* American Mathematical Society Bookstore, (p 63). ISBN 0821844032.

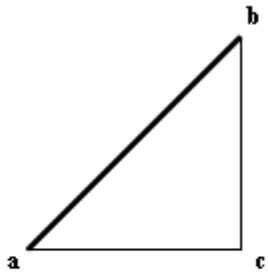
Sireci, S.G., Robin, F., & Patellis, T. (1997). Using cluster analysis to facilitate the standard-setting process. Paper presented at the *103rd convention of the National Council of Measurement in Education*, Chicago, USA.

Violato, C., Marini, A., Lee, C., (2003). A validity study of expert judgment procedures for setting cutoff scores on high stakes credentialing examinations using cluster analysis. *Evaluation & The Health Professions*, 26(1). 59-72.

Wilson, N. (1998). Educational Standards and the Problem of Error. *Education Policy Analysis Archives*, 6(10).

Subjects/disciplines	Core knowledge test items	Desirable test items
Anatomy	10	6
Biochemistry	5	5
Physiology	15	7
Pharmacology	10	6
Pathology	10	6
Community med.	4	2
Clinical areas	10	4
TOTAL Test Items	**64** +	**36** = **100**

Table 1: Break-up of MCQ test items in a sample test (Hypothetical)



$ab^2 = ac^2 + cb^2$

Figure 1: Calculation of the value of the hypotenuse using the Pythagorean theorem

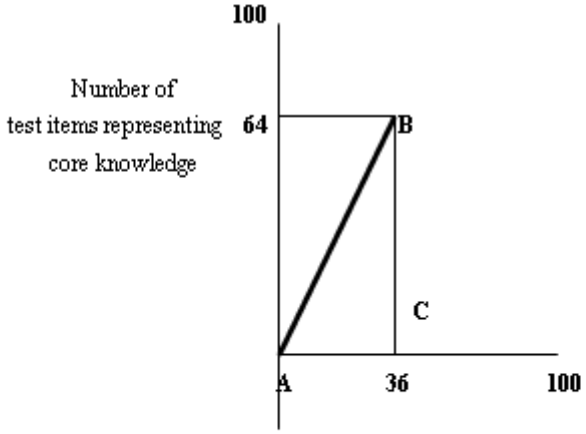


Figure 2: Relationship between test items representing core knowledge and test items representing desirable knowledge in a given test with 100 test items (Hypothetical).

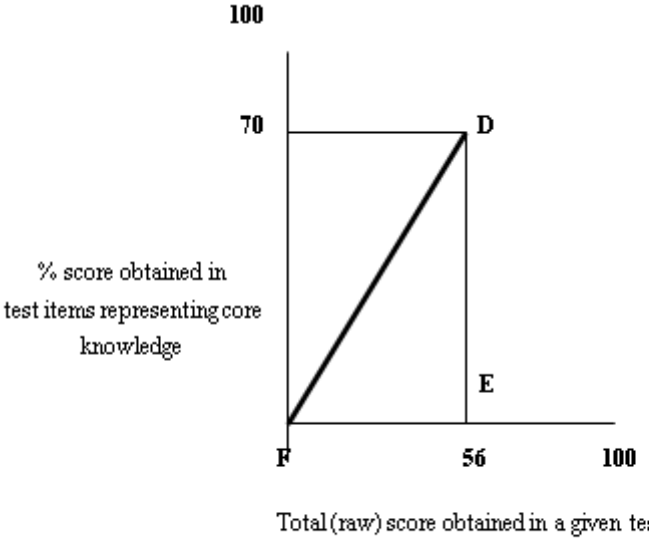


Figure 3: Relationship between a student's % score obtained in test items representing core knowledge and the Total (raw) score obtained in the test. (Hypothetical)